Variational Bayes Solution of Linear Neural Networks and its Generalization Performance

Shinichi Nakajima [†‡] and Sumio Watanabe [†]
[†] Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuda, Midori-ku,
Yokohama, Kanagawa, 226-8503 Japan
tel. & fax. +81-45-924-5018
nakajima.s@cs.pi.titech.ac.jp, swatanab@pi.titech.ac.jp
[‡] Nikon Corporation
201-9 Oaza-Miizugahara,
Kumagaya, Saitama, 360-8559 Japan

February 24, 2007

Abstract

It is well-known that, in unidentifiable models, the Bayes estimation provides much better generalization performance than the maximum likelihood (ML) estimation. However, its accurate approximation by Markov chain Monte Carlo methods requires huge computational costs. As an alternative, a tractable approximation method, called the variational Bayes (VB) approach, has recently been proposed and been attracting people's attention. Its advantage over the expectation maximization (EM) algorithm, often used for realizing the ML estimation, has been experimentally shown in many applications, nevertheless, has not been theoretically shown yet. In this paper, through the analysis of the simplest unidentifiable models, we theoretically show some properties of the VB approach. We first prove that, in three-layer linear neural networks, the VB approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation. Then, we theoretically clarify its free energy, generalization error, and training error. Comparing them with those of the ML estimation and of the Bayes estimation, we discuss the advantage of the VB approach. We also show that, unlike in the Bayes estimation, the free energy and the generalization error are less simply related with each other, and that, in typical cases, the VB free energy well approximates the Bayes one, while the VB generalization error significantly differs from the Bayes one.

1 Introduction

The conventional learning theory [Cramer, 1949], which provides the mathematical foundation of information criteria, such as Akaike's information criterion (AIC) [Akaike, 1974], Bayesian information criterion (BIC) [Schwarz, 1978], the minimum description length criterion (MDL) [Rissanen, 1986], etc., describes the asymptotic behavior of the Bayes free energy, the generalization error, the training error, etc., in the regular statistical models. However, many models used in the field of machine learning, such as neural networks, Bayesian networks, mixture models, hidden Markov models, etc., are, actually, classified as unidentifiable models, which have singularities in the parameter spaces. Around the singularities, on which the Fisher information matrix is singular, the log-likelihood cannot be approximated by any quadratic form of the parameter. Therefore, neither the distribution of the maximum likelihood estimator nor the Bayes posterior distribution asymptotically converges to the normal distribution, which prevents the conventional learning theory to hold [Hartigan, 1985; Watanabe, 1995; Amari *et al.*, 2006; Hagiwara, 2002]. Accordingly, the information criteria have no theoretical foundation in unidentifiable models.

Some properties of learning in unidentifiable models have been theoretically clarified. In the maximum likelihood (ML) estimation, which is asymptotically equivalent to the maximum a posterior (MAP) estimation, the asymptotic behavior of the log-likelihood ratio in some unidentifiable models was analyzed [Hartigan, 1985; Bickel and Chernoff, 1993; Takemura and Kuriki, 1997; Kuriki and Takemura, 2001; Fukumizu, 2003], with facilitation by the idea of the locally conic parameterization [Dacunha-Castelle and Gassiat, 1997]. It has, thus, been known that the ML estimation, in general, provides poor generalization performance, and in the worst cases, the ML estimator diverges. In linear neural networks, on which we focus in this paper, the generalization error was clarified and proved to be greater than that of the regular models whose dimension of the parameter space is the same when the model is redundant to learn the true distribution [Fukumizu, 1999], although the ML estimator is in a finite region [Baldi and Hornik, 1995].

On the other hand, for the analysis of the Bayes estimation in unidentifiable models, an algebraic geometrical method was developed [Watanabe, 2001a], by which the generalization errors or their upper bounds in some unidentifiable models were clarified and proved to be less than that of the regular models [Watanabe, 2001a; Yamazaki and Watanabe, 2002; Rusakov and Geiger, 2002; Yamazaki and Watanabe, 2003a; 2003b; Aoyagi and Watanabe, 2005], and moreover, the generalization error of any model having singularities was proved to be less than that of the regular models when we use a prior distribution having positive values on the singularities [Watanabe, 2001b].

According to the previous works above, it can be said that, in unidentifiable models, the Bayes estimation has the advantage of generalization performance over the ML estimation. However, the Bayes posterior distribution can seldom be exactly realized. Furthermore, Markov chain Monte Carlo (MCMC) methods, often used for approximation of the posterior distribution, require huge computational costs. As an alternative, the variational Bayes (VB) approach, which provides computational tractability, was proposed [Hinton and van Camp, 1993; MacKay, 1995; Attias, 1999; Jaakkola and Jordan, 2000; Ghahramani and Beal, 2001]. In models with latent variables, an iterative algorithm like the expectation maximization (EM) algorithm, which is to provide the ML estimator [Dempster *et al.*, 1977], was derived in the framework of the VB approach [Attias, 1999; Ghahramani and Beal, 2001]. Some properties similar to the EM algorithm has been proved [Wang and Titterington, 2004], and it has been shown that the variational Bayes approach can be considered as the natural gradient descent with respect to the VB free energy [Sato, 2001]. However, its advantage of generalization performance over the EM algorithm has not been theoretically shown yet, while it has been experimentally shown in many applications. In addition, although the VB free energies in some models have been clarified [Watanabe and Watanabe, 2006; Hosino *et al.*, 2005; Nakano and Watanabe, 2005], they, unfortunately, provide little information on their generalization errors unlike in the Bayes estimation. It is because the simple relation that holds in the Bayes estimation between the free energy and the generalization error[1] does not hold in the VB approach. Currently, in any unidentifiable model, the VB generalization error has never been theoretically clarified yet.

By the way, it was, surprisingly, proved that, even in the regular models, the usual ML estimator does not provide the optimum generalization performance [Stein, 1956]. The James-Stein (JS) shrinkage estimator, which dominates the ML estimator, was subsequently introduced [James and Stein, 1961].[2] Some papers discussed the relation between the JS estimator and Bayesian learning methods as follows: it was shown that the JS estimator can be derived as the solution of an empirical Bayes (EB) approach where the prior distribution of a regular model has a hyperparameter as its variance to be estimated based on the marginal likelihood [Efron and Morris, 1973]; the similarity between the asymptotic behavior of the generalization error of the Bayes estimation in a simple class of unidentifiable models and that of the JS estimation in the regular models was pointed out [Watanabe and Amari, 2003]; it was proved that, in three-layer linear neural networks, a subspace Bayes (SB) approach, the extension of the EB approach where a part of the parameters are regarded as hyperparameters, is asymptotically equivalent to a positive-part JS type shrinkage estimation, a subspecies of the JS estimation [Nakajima and Watanabe, 2005b; 2006]. Interestingly, the VB solution, derived in this paper, is similar to the SB solution.

In this paper, we consider the theoretical reason why the VB approach provides good generalization performance, through the analysis of the simplest unidentifiable models. We first derive the VB solution of three-layer linear neural networks, which reveals an interesting relation between the VB approach, a rising method in the field of machine learning, and the JS shrinkage estimation, a classic in statistics. Then, we clarify its free energy, generalization error, and training error. The major results of this paper are as follows:

1. The VB approach is asymptotically equivalent to a positive-part JS type shrinkage estimation.

   Hence, the VB estimator of a necessary component to realize the true distribution is asymptotically equivalent to the ML estimator; while the VB estimator of a redundant component is not equivalent to the ML estimator even in the asymptotic limit.

[1] This relation will appear as Eq.(21) in Section 3.

[2] The definition of the verb ”dominate” will be given in Section 3.3.

2. In the VB approach, the asymptotic behavior of the free energy and that of the generalization error are less simply related to each other, unlike in the Bayes estimation.

   In typical cases, the VB free energy well approximates the Bayes one; while the VB generalization error significantly differs from the Bayes one. Moreover, their dependences on the redundant parameter dimension are different from each other, which can throw a doubt on the model selection by minimizing the VB free energy to obtain better generalization performance.

3. Although the VB free energy is, by definition, never less than the Bayes one, the VB generalization error can be much less than the Bayes one.

   This, however, does not mean the domination of the VB approach over the Bayes estimation, and is consistent with the proved superiority of the Bayes estimation.

4. The more different the parameter space dimensions of the different layers are, the smaller the generalization error is.

   This implies that the advantage of the VB approach over the EM algorithm can be enhanced in mixture models, where the dimension of the upper layer parameters, which corresponds to the number of the mixing coefficients, is one per component and the dimension of the lower layer parameters, which corresponds to the number of the parameters that each component has, is usually more.

5. The VB approach has not only a property similar to the Bayes estimation but also another property similar to the ML estimation.

In Section 2, neural networks and linear neural networks are described. The Bayes estimation, the VB approach, and the JS type shrinkage estimator are introduced in Section 3. Then, the effects of singularities upon generalization performance, and the significance of consideration of *delicate* situations are explained in Sections 4 and 5, respectively. After that, the solution of the VB predictive distribution is derived in Section 6, and then its free energy, generalization error, and training error are clarified in Section 7. In Section 8, the theoretical values of the generalization properties of the VB approach are compared with those of the ML estimation and of the Bayes estimation. Discussion and conclusions follow in Sections 9 and 10, respectively.

2 Linear Neural Networks

Let $x \in \mathbb{R}^M$ be an input (column) vector, $y \in \mathbb{R}^N$ an output vector, and w a parameter vector. A neural network model can be described as a parametric family of maps $\{f(\cdot; w) : \mathbb{R}^M \mapsto \mathbb{R}^N\}$. A three-layer neural network with H hidden units is defined by

$$f(x; w) = \textstyle\sum_{h=1}^{H} b_h \psi\left(a_h^t x\right), \tag{1}$$

where $w = \{(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N; h = 1, \ldots, H\}$ summarizes all the parameters, $\psi(\cdot)$ is an activation function, which is usually a bounded, non-decreasing, antisymmetric, nonlinear function like $\tanh(\cdot)$, and t denotes the transpose of a matrix or vector. We denote by $\mathcal{N}_d(\mu, \varSigma)$ the d-dimensional normal distribution with average vector μ and covariance matrix $\varSigma$, and by $\mathcal{N}_d(\cdot; \mu, \varSigma)$ its density function. Assume that the output is observed with a noise subject to $\mathcal{N}_N(0, \varSigma)$. Then, the conditional distribution is given by

$$p(y|x, w) = \mathcal{N}_N(y; f(x; w), \varSigma). \tag{2}$$

In this paper, we focus on linear neural networks, whose activation functions are linear, as the simplest unidentifiable models. A linear neural network model (LNN) is defined by

$$f(x; A, B) = BAx, \tag{3}$$

where $A = (a_1, \ldots, a_H)^t$ is an $H \times M$ input parameter matrix and $B = (b_1, \ldots, b_H)$ is an $N \times H$ output parameter matrix. Because the transform $(A, B) \mapsto (TA, BT^{-1})$ does not change the map for any non-singular $H \times H$ matrix T, the parameterization in Eq.(3) has trivial redundancy. Accordingly, the *essential* dimension of the parameter space is given by

$$K = H(M + N) - H^2 = H(L + l) - H^2, \tag{4}$$

where

$$L = \max(M, N), \tag{5}$$
$$l = \min(M, N). \tag{6}$$

We assume that $H \leq l$ throughout this paper. An LNN, also known as a reduced-rank regression model, is not a toy but an useful model in many applications [Reinsel and Velu, 1998]. It is used for multivariate factor analysis when the relevant dimension of the factors is unknown, and classified as an *essentially* singular model when $0 < H < l$ because no transform makes it regular. Although LNNs are simple, we expect that some phenomena caused by the singularities, revealed in this paper, would be observed also in other unidentifiable models.

3 Learning Methods

3.1 Bayes Estimation

Let $X^n = \{x_1, \ldots, x_n\}$ and $Y^n = \{y_1, \ldots, y_n\}$ be arbitrary n training samples independently and identically taken from the true distribution $q(x, y) = q(x)q(y|x)$. The marginal conditional likelihood of a model $p(y|x, w)$ is given by

$$Z(Y^n|X^n) = \int \phi(w) \textstyle\prod_{i=1}^n p(y_i|x_i, w) dw, \tag{7}$$

where $\phi(w)$ is a prior distribution. The posterior distribution is given by

$$p(w|X^n, Y^n) = \frac{\phi(w) \prod_{i=1}^n p(y_i|x_i, w)}{Z(Y^n|X^n)}. \tag{8}$$

The predictive distribution is defined as the average of the model over the posterior distribution:

$$p(y|x, X^n, Y^n) = \int p(y|x, w)p(w|X^n, Y^n)dw. \tag{9}$$

The free energy, evidence, or stochastic complexity, which is used for model selection or hyperparameter estimation [Efron and Morris, 1973; Akaike, 1980; MacKay, 1992], is defined by

$$F(Y^n|X^n) = -\log Z(Y^n|X^n). \tag{10}$$

Commonly, the normalized free energy, defined by

$$F(n) = \langle F(Y^n|X^n) + \log q(Y^n|X^n) \rangle_{q(X^n,Y^n)}, \tag{11}$$

where $\langle \cdot \rangle_{q(X^n,Y^n)}$ denotes the expectation value over all sets of n training samples, can be asymptotically expanded as follows:

$$F(n) = \lambda' \log n + o(\log n), \tag{12}$$

where λ' is called the free energy coefficient in this paper. The generalization error, a criterion of generalization performance, and the training error are defined by

$$G(n) = \langle G(X^n, Y^n) \rangle_{q(X^n,Y^n)}, \tag{13}$$
$$T(n) = \langle T(X^n, Y^n) \rangle_{q(X^n,Y^n)}, \tag{14}$$

respectively, where

$$G(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dxdy \tag{15}$$

is the Kullback-Leibler (KL) divergence of the predictive distribution from the true distribution, and

$$T(X^n, Y^n) = n^{-1} \textstyle\sum_{i=1}^n \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} \tag{16}$$

is the empirical KL divergence. Commonly, the generalization error and the training error can be asymptotically expanded as follows:

$$G(n) = \lambda n^{-1} + o(n^{-1}), \tag{17}$$
$$T(n) = \nu n^{-1} + o(n^{-1}), \tag{18}$$

where the coefficients of the leading terms, λ and ν, are called the generalization coefficient and the training coefficient, respectively, in this paper.

In the regular models, the free energy, the generalization, and the training coefficients are simply related to the parameter dimension K in the Bayes estimation, as well as in the ML estimation, as follows:

$$2\lambda' = 2\lambda = -2\nu = K \tag{19}$$

The penalty factor of AIC [Akaike, 1974] is based on the fact that $(\lambda - \nu) = K$, and the penalty factor of BIC [Schwarz, 1978], as well as that of the MDL [Rissanen, 1986], is based on the fact that $2\lambda' = K$. However, in unidentifiable models, Eq.(19) does not generally hold and the coefficients can depend not only on the parameter dimension of the learning machine but also on the true distribution. So, it seems to be difficult to propose any information criterion in unidentifiable models, nevertheless, an information criterion, called a singular information criterion, has recently been proposed rather by utilizing the dependence on the true distribution [Yamazaki *et al.*, 2005]. For such kind of work, clarifying the coefficients is important.

In the (rigorous) Bayes estimation, we can prove that the relation

$$\lambda' = \lambda \tag{20}$$

still holds even in unidentifiable models, by using the following well-known relation [Levin *et al.*, 1990]:

$$G(n) = F(n+1) - F(n). \tag{21}$$

Therefore, clarifying the Bayes free energy immediately informs us of the asymptotic behavior of the Bayes generalization error.

3.2 Variational Bayes Approach

For an arbitrary trial distribution $r(w)$, Jensen's inequality leads to the following one:

$$F(Y^n|X^n) \leq \left\langle \log \frac{r(w)}{p(Y^n|X^n, w)\phi(w)} \right\rangle_{r(w)} = \bar{F}(Y^n|X^n), \tag{22}$$

where $\langle \cdot \rangle_p$ denotes the expectation value over a distribution p, and $\bar{F}(Y^n|X^n)$ is called the generalized free energy in this paper. In the VB approach, restricting the set of possible distributions for $r(w)$, we minimize the generalized free energy, a functional of $r(w)$. The optimum of the trial distribution, which we denote by $\hat{r}(w)$, is called the VB posterior distribution, and substituted for $p(w|X^n, Y^n)$ in Eq.(9). In addition, the minimum value of the generalized free energy, which we denote by $\tilde{F}(Y^n|X^n)$, is called the VB free energy, and the expectation value over the VB posterior distribution is called the VB estimator.

In the original VB approach, proposed for learning of neural networks, the posterior distribution is restricted to the normal distribution [Hinton and van Camp, 1993]. Recently, an iterative algorithm with good tractability has been proposed for models with latent variables, such as mixture models, graphical models, etc., only by using an appropriate class of prior distributions and restricting the posterior distribution such that the parameters and the latent variables are independent of each other, and since been attracting people's attention [Attias, 1999; Ghahramani and Beal, 2001].

In LNNs, we can deduce a similar algorithm by restricting the posterior distribution such that the parameters of different layers, as well as those of different components, are independent

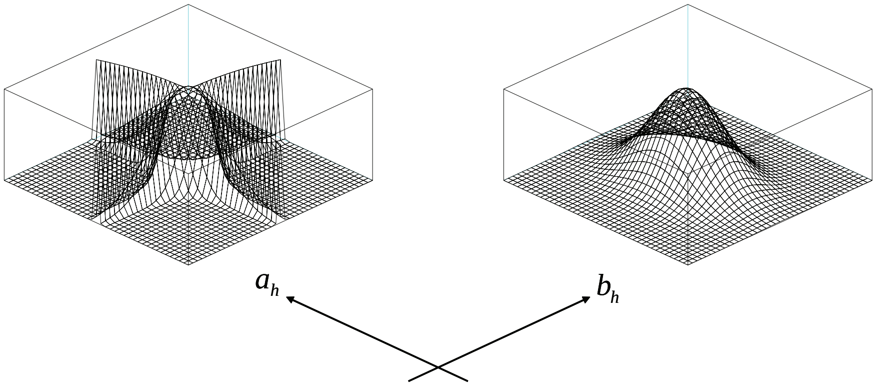

Figure 1: VB posterior distribution (right-hand side), which is the independent distribution with respect to different layer parameters to be substituted for the Bayes posterior distribution (left-hand side).

of each other, as shown below.[3] Assume that the VB posterior distribution factorizes as

$$r(w) = r(A, B) = r(A)r(B). \tag{23}$$

Then, the generalized free energy is given by

$$\bar{F}(Y^n|X^n) = \int r(A)r(B) \log \frac{r(A)r(B)}{p(Y^n|X^n, A, B)\phi(A, B)} dAdB, \tag{24}$$

where $\int dV$ denotes the integral with respect to all the elements of the matrix V. Using the variational method, we can easily show that the VB posterior distribution satisfies the following relations, if we use a prior distribution that factorizes as $\phi(w) = \phi(A, B) = \phi(A)\phi(B)$:

$$r(A) \propto \phi(A) \exp\langle \log p(Y^n|X^n, A, B)\rangle_{r(B)}, \tag{25}$$

$$r(B) \propto \phi(B) \exp\langle \log p(Y^n|X^n, A, B)\rangle_{r(A)}. \tag{26}$$

We find from Eqs.(25) and (26) that the VB posterior distributions are the normal if the prior distributions, $\phi(A)$ and $\phi(B)$, are the normal, because the log-likelihood of an LNN, $\log p(Y^n|X^n, A, B)$, is a biquadratic function of A and B. Figure 1 illustrates the VB posterior distribution (right-hand side). The VB posterior distribution, which is the independent distribution with respect to different layer parameters minimizing the generalized free energy, Eq.(24), is substituted for the Bayes posterior distribution (left-hand side). We then apply another restriction, the independence of the parameters of different components, to the posterior distribution:

$$r(A, B) = \textstyle\prod_{h=1}^{H} r(a_h)r(b_h). \tag{27}$$

[3] Note that this restriction is similar to the one applied in models with hidden variables in [Attias, 1999]. Therefore, the analysis in this paper also provides an insight into the properties of the VB approach in more general unidentifiable models, which will be discussed in Section 9.4.

This restriction makes the relations, Eqs.(25) and (26), decompose into those for each component as follows, if we use a prior distribution that factorizes as $\phi(A, B) = \prod_{h=1}^{H} \phi(a_h)\phi(b_h)$:

$$r(a_h) \propto \phi(a_h) \exp\langle \log p(Y^n|X^n, A, B)\rangle_{r(A,B)/r(a_h)}, \tag{28}$$

$$r(b_h) \propto \phi(b_h) \exp\langle \log p(Y^n|X^n, A, B)\rangle_{r(A,B)/r(b_h)}, \tag{29}$$

where $\langle\cdot\rangle_{r(A,B)/r(a_h)}$, as well as $\langle\cdot\rangle_{r(A,B)/r(b_h)}$, denotes the expectation value over the trial distribution of the parameters except a_h, as well as that except b_h.

In the same way as the Bayes estimation, the predictive distribution, the normalized free energy, the generalization error, and the training error of the VB approach are given by

$$\tilde{p}(y|x, X^n, Y^n) = \int p(y|x, w)\hat{r}(w)dw, \tag{30}$$

$$\tilde{F}(n) = \langle \tilde{F}(Y^n|X^n) + \log q(Y^n|X^n)\rangle_{q(X^n,Y^n)} = \tilde{\lambda}' \log n + o(\log n), \tag{31}$$

$$\tilde{G}(n) = \langle \tilde{G}(X^n, Y^n)\rangle_{q(X^n,Y^n)} = \tilde{\lambda} n^{-1} + o(n^{-1}), \tag{32}$$

$$\tilde{T}(n) = \langle \tilde{T}(X^n, Y^n)\rangle_{q(X^n,Y^n)} = \tilde{\nu} n^{-1} + o(n^{-1}), \tag{33}$$

respectively, where

$$\tilde{G}(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{\tilde{p}(y|x, X^n, Y^n)} dxdy, \tag{34}$$

$$\tilde{T}(X^n, Y^n) = n^{-1} \textstyle\sum_{i=1}^{n} \log \frac{q(y|x)}{\tilde{p}(y|x, X^n, Y^n)}. \tag{35}$$

In general, $\tilde{\lambda}' \neq \tilde{\lambda}$, because the relation corresponding to Eq.(21) does not hold in the VB approach. Accordingly, the VB free energy and the VB generalization error are less simply related to each other, as will be shown in Section 8.

3.3 James-Stein type Shrinkage Estimator

In this subsection, considering a regular, i.e., identifiable, model, we introduce the James-Stein (JS) type shrinkage estimator. Suppose that $Z^n = \{z_1, \ldots, z_n\}$ are arbitrary n training samples independently and identically taken from $\mathcal{N}_d(\mu, I_d)$, where μ is the mean parameter to be estimated, and I_d denotes the $d \times d$ identity matrix. We denote by $*$ the true value of a parameter and by a *hat* an estimator of a parameter throughout this paper. We define the risk as the mean squared error, i.e., $g_{\hat{\mu}}(\mu^*) = \langle \|\hat{\mu} - \mu^*\|^2 \rangle_{q(Z^n)}$.[4] We say that $\hat{\mu}_\alpha$ dominates $\hat{\mu}_\beta$ if $g_{\hat{\mu}_\alpha}(\mu^*) \leq g_{\hat{\mu}_\beta}(\mu^*)$ for arbitrary μ^* and $g_{\hat{\mu}_\alpha}(\mu^*) < g_{\hat{\mu}_\beta}(\mu^*)$ for a certain μ^*. The usual ML estimator, $\hat{\mu}_{MLE} = n^{-1} \sum_{i=1}^{n} z_i$, is, naturally, an efficient estimator, which is never dominated by any unbiased estimator. However, the ML estimator has, surprisingly, been proved to be inadmissible when $d \geq 3$, i.e., there exists at least one biased estimator that dominates the

[4] The risk function coincides with the twice of the generalization error, Eq.(13), when $p(z|\hat{\mu}) = \mathcal{N}_d(z; \hat{\mu}, I_d)$ is regarded as the predictive distribution.

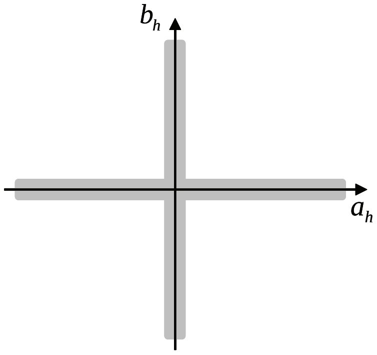

Figure 2: Singularities of a neural network model.

ML estimator [Stein, 1956]. The JS shrinkage estimator, defined as follows, was, subsequently, introduced. [James and Stein, 1961]:

$$\hat{\mu}_{JS} = (1 - \chi/n\|\hat{\mu}_{MLE}\|^2)\hat{\mu}_{MLE}, \tag{36}$$

where $\chi = (d-2)$ is called the degree of shrinkage in this paper. The estimators expressed by Eq.(36) with arbitrary $\chi > 0$ are called the JS type shrinkage estimators. We can easily find that a positive-part JS type shrinkage estimator, defined by

$$\hat{\mu}_{PJS} = \theta(n\|\hat{\mu}_{MLE}\|^2 > \chi)(1 - \chi/n\|\hat{\mu}_{MLE}\|^2)\hat{\mu}_{MLE} = \left(1 - \chi\chi'^{-1}\right)\hat{\mu}_{MLE}, \tag{37}$$

dominates the JS type shrinkage estimator, Eq.(36), with the same degree of shrinkage, χ. Here, $\theta(\cdot)$ is the indicator function of an event, i.e., which is equal to one if the event is true and to zero otherwise, and $\chi' = \max(\chi, n\|\hat{\mu}_{MLE}\|^2)$. Note that, even in the large sample limit, the shrinkage estimators are not equivalent to the ML estimator if $\mu^* = 0$, although they are equivalent to the ML estimator otherwise. Also note that the indicator function in Eq.(37) works like a model selection method, which eliminates insignificant parameters, and the shrinkage factor works like a regularizer, which pulls the estimator near the model with smaller degree of freedom.

4 Unidentifiability and Singularities

We say that a parametric model is unidentifiable if the map from the parameter to the probability distribution is not one-to-one. A neural network model, Eq.(1), is unidentifiable because the model is invariant for any a_h if $b_h = 0$, or vice versa. The continuous points denoting the same distribution are called the singularities, because the Fisher information matrix on them is singular. The shadowed locations in Fig. 2 indicate the singularities. We can see in Fig. 2 that the model denoted by the singularities has more neighborhoods and state density than any other model denoted by only one point each. When the true model is not on the singularities, they asymptotically do not affect prediction, and therefore, the conventional learning theory of the regular models holds. On the other hand, when the true model is on the singularities, they significantly affect generalization performance as follows:

***ML type effect*:** In the ML estimation, as well as the MAP estimation, increase of the neighborhoods of the true distribution leads to increase of the flexibility of imitating noises, and therefore, accelerates overfitting.

***Bayesian effect*:** In the Bayesian learning methods, the large state density of the true distribution increases its weight, and therefore, suppresses overfitting.

According to the effects above, we can expect the Bayesian learning methods to provide better generalization performance than the ML estimation.

5 Significance of *Delicate* Situation Analysis

Suppression of overfitting accompanies insensitivity to the true components with small amplitude. There is a trade-off, which would, however, be ignored in asymptotic analysis if we would consider only situations when the true model is *distinctly* on the singularities or not. Consider, for example, the positive-part JS type estimator, Eq.(37), with a very large but finite degree of shrinkage, $1 \ll \chi < \infty$. That estimator provides very small generalization error when $\mu^* = 0$ because of its strong shrinkage, and the same generalization error as that of the ML estimator when $\mu^* = O(1)$ in the asymptotic limit because $\hat{\mu}_{PJS} = \hat{\mu}_{MLE} + O_p(n^{-1})$. Therefore, in ordinary asymptotic analysis, it could seem to be a good learning method. However, such an estimator will not perform well, because it can provide extremely worse generalization performance in the real situations where the number of training samples, n, is finite. In other words, the ordinary asymptotic analysis ignores the insensitivity to the true small components.

To consider the case where n is sufficiently large but finite, with the benefit of the asymptotic approximation, we should analyze the *delicate* situations [Watanabe and Amari, 2003] when $0 < \sqrt{n}\|\mu^*\| < \infty$. We have to balance between the generalization performance in the region where $\|\mu^*\| = 0$ and that in the region where $0 < \sqrt{n}\|\mu^*\| < \infty$, when we select a learning method. The *delicate* situations in unidentifiable models should be defined as the ones where the KL divergence of the true distribution from the singularities is comparable to n^{-1}. This kind of analysis is important in model selection problems and in statistical tests with finite number of samples for the following reasons: first, that there naturally exist a few true components with amplitude comparable to $n^{-1/2}$ when neither the smallest nor the largest model is selected; and secondly, that whether the selected model involves such components essentially affects generalization performance. In this paper, we discuss the generalization error and the training error in the *delicate* situations, in addition to the ordinary asymptotic analysis.

6 Theoretical Analysis

6.1 Variational Condition

Assume that the covariance matrix of a noise is known and equal to I_N. Then, the conditional distribution of an LNN is given by

$$p(y|x, A, B) = \mathcal{N}_N(y; BAx, I_N). \tag{38}$$

We use the prior distribution that factorizes as $\phi(A, B) = \phi(A)\phi(B)$, where

$$\phi(A) = \mathcal{N}_{HM}(\mathrm{vec}(A^t); 0, c_a^2 I_{HM}), \tag{39}$$

$$\phi(B) = \mathcal{N}_{NH}(\mathrm{vec}(B); 0, c_b^2 I_{NH}). \tag{40}$$

Here $0 < c_a^2, c_b^2 < \infty$ are constants corresponding to the variances, and vec$(\cdot)$ denotes the vector created from a matrix by stacking the column vectors below one another. Assume that the true conditional distribution is $p(y|x, A^*, B^*)$, where B^*A^* is the true map with rank $H^* \leq H$. For simplicity, we assume that the input vector is orthonormalized so that $\int xx^t q(x)dx = I_M$. Consequently, the central limit theorem leads to the following two equations with respect to the sufficient statistics:

$$Q(X^n) = n^{-1}\textstyle\sum_{i=1}^n x_i x_i^t = I_M + O_p(n^{-1/2}), \tag{41}$$

$$R(X^n, Y^n) = n^{-1}\textstyle\sum_{i=1}^n y_i x_i^t = B^*A^* + O_p(n^{-1/2}), \tag{42}$$

where $Q(X^n)$ is an $M \times M$ symmetric matrix and $R(X^n, Y^n)$ is an $N \times M$ matrix. Hereafter, we abbreviate $Q(X^n)$ as Q and $R(X^n, Y^n)$ as R.

Let γ_h be the h-th largest singular value of the matrix $RQ^{-1/2}$, ω_{a_h} the corresponding right singular vector, and ω_{b_h} the corresponding left singular vector, where $1 \leq h \leq H$. We find from Eq.(42) that, in the asymptotic limit, the singular values corresponding to the necessary components to realize the true distribution converge to finite values; while the others corresponding to the redundant components converge to zero. Therefore, with probability 1, the largest H^* singular values correspond to the necessary components, and the others correspond to the redundant components. Combining that with Eq.(41), we have

$$\omega_{b_h} RQ^{\rho} = \omega_{b_h} R + O_p(n^{-1}) \qquad \text{for} \qquad H^* < h \leq H, \tag{43}$$

where $-\infty < \rho < \infty$ is an arbitrary constant.

Substituting Eqs.(38)–(40) into Eqs.(28) and (29), we find that the posterior distribution can be written as follows:

$$r(A, B) = \textstyle\prod_{h=1}^H r(a_h) r(b_h), \tag{44}$$

where

$$r(a_h) = \mathcal{N}_M(a_h; \mu_{a_h}, \Sigma_{a_h}), \tag{45}$$

$$r(b_h) = \mathcal{N}_N(b_h; \mu_{b_h}, \Sigma_{b_h}). \tag{46}$$

Given an arbitrary map BA, we can have A with its orthogonal row vectors and B with its orthogonal column vectors by using the singular value decomposition. Just in that case, both of the prior probabilities, Eqs.(39) and (40), are maximized. Accordingly, $\{\mu_{a_h}; h = 1, \ldots, H\}$, as well as $\{\mu_{b_h}; h = 1, \ldots, H\}$, of the optimum distribution is a set of vectors orthogonal to each other.

Now, we can easily obtain the following condition, called the variational condition, by substituting Eqs.(44)–(46) into Eqs.(28) and (29):

$$\mu_{a_h} = n\Sigma_{a_h} R^t \mu_{b_h}, \tag{47}$$

$$\Sigma_{a_h} = (n(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h}))Q + c_a^{-2} I_M)^{-1}, \tag{48}$$

$$\mu_{b_h} = n\Sigma_{b_h} R \mu_{a_h}, \tag{49}$$

$$\Sigma_{b_h} = (n(\mu_{a_h}^t Q \mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})) + c_b^{-2})^{-1} I_N, \tag{50}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Similarly, by substituting Eqs.(44)–(46) into Eq.(24), we also obtain the following form of the generalized free energy:

$$2\bar{F}(Y^n|X^n) = \sum_{h=1}^H \Big\{ -\log \sigma_{a_h}^{2M} \sigma_{b_h}^{2N} + \frac{\|\mu_{a_h}\|^2 + M\sigma_{a_h}^2}{c_a^2} + \frac{\|\mu_{b_h}\|^2 + N\sigma_{b_h}^2}{c_b^2} -2n\mu_{b_h}^t R \mu_{a_h} + n\big(\mu_{a_h}^t Q \mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})\big)\big(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})\big) \Big\} + \sum_{i=1}^n \|y_i\|^2 + \mathrm{const.}, \tag{51}$$

where $\sigma_{a_h}^2$ and $\sigma_{b_h}^2$ are defined by

$$\Sigma_{a_h} = \sigma_{a_h}^2 (I_M + O_p(n^{-1/2})), \tag{52}$$

$$\Sigma_{b_h} = \sigma_{b_h}^2 (1 + O_p(n^{-1/2})) I_N, \tag{53}$$

respectively.

6.2 Variational Bayes Solution

The variational condition, Eqs.(47)–(50), can be analytically solved, which leads to the following theorem:

Theorem 1 *Let $L'_h = \max(L, n\gamma_h^2)$. The VB estimator of the map of an LNN is given by*

$$\hat{B}\hat{A} = \sum_{h=1}^H \left(1 - L L_h'^{-1}\right) \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \tag{54}$$

(The proof is given in Appendix A.)

Moreover, we obtain the following lemma, expanding the predictive distribution divided by the true distribution, i.e., $\tilde{p}(y|x, X^n, Y^n)/q(y|x)$:

Lemma 1 *The predictive distribution of an LNN in the VB approach can be written as follows:*

$$\tilde{p}(y|x, X^n, Y^n) = \mathcal{N}_N(y; \hat{V}\hat{B}\hat{A}x, \hat{V}) + O_p(n^{-3/2}), \tag{55}$$

where $\hat{V} = I_N + O_p(n^{-1})$.

Comparing Eq.(54) with the ML estimator

$$\hat{B}\hat{A}_{MLE} = \textstyle\sum_{h=1}^{H} \omega_{b_h} \omega_{b_h}^t R Q^{-1} \tag{56}$$

[Baldi and Hornik, 1995], we find that the VB estimator of each component is asymptotically equivalent to a positive-part JS type shrinkage estimator, defined by Eq.(37). However, even if the VB estimator is asymptotically equivalent to the shrinkage estimator, the VB approach can differ from the shrinkage estimation because of the extent of the VB posterior distribution. Lemma 1 states that the VB posterior distribution is sufficiently localized, so that we can substitute the model at the VB estimator for the VB predictive distribution with asymptotically insignificant impact upon generalization performance. Therefore, we conclude that the VB approach is asymptotically equivalent to the shrinkage estimation. Note that the variances, c_a^2 and c_b^2, of the prior distributions, Eqs.(39) and (40), asymptotically have no effect upon prediction and hence upon generalization performance, as far as they are positive and finite constants.

7 Generalization Properties

7.1 Free Energy

We can obtain the VB free energy by substituting the VB posterior distribution, derived in Appendix A, into Eq.(51), where only the order of $\hat{\sigma}_{a_h}^2$ and that of $\hat{\sigma}_{b_h}^2$ signify because the leading term of the normalized free energy should be the first one in the curly braces of Eq.(51).

Theorem 2 *The normalized free energy of an LNN in the VB approach can be asymptotically expanded as*

$$\tilde{F}(n) = \tilde{\lambda}' \log n + O(1),$$

where the free energy coefficient is given by

$$2\tilde{\lambda}' = H^*(L + l) + (H - H^*)l. \tag{57}$$

(Proof) We will separately consider the necessary components, which imitate the true ones with positive singular values, and the redundant ones. For a necessary component, $h \leq H^*$, Eq.(84) implies that both $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are of order $O_p(1)$. Then, we find from Eqs.(81) and (83) that both Σ_{a_h} and Σ_{b_h} are of order $O_p(n^{-1})$. Substituting the fact above into Eq.(51) leads to the first term of Eq.(57). For a redundant component, $h > H^*$, we find from the VB solution, Eqs.(128)–(139), that the orders of the estimators are as follows:

1. When $M > N$,

$$\hat{\mu}_{a_h} = O_p(1), \qquad \hat{\sigma}_{a_h}^2 = O_p(1), \qquad \hat{\mu}_{b_h} = O_p(n^{-1/2}), \qquad \hat{\sigma}_{b_h}^2 = O_p(n^{-1}). \tag{58}$$

2. When $M = N$,

$$\hat{\mu}_{a_h} = O_p(n^{-1/4}), \quad \hat{\sigma}_{a_h}^2 = O_p(n^{-1/2}), \quad \hat{\mu}_{b_h} = O_p(n^{-1/4}), \quad \hat{\sigma}_{b_h}^2 = O_p(n^{-1/2}). \tag{59}$$

3. When $M < N$,

$$\hat{\mu}_{a_h} = O_p(n^{-1/2}), \qquad \hat{\sigma}^2_{a_h} = O_p(n^{-1}), \qquad \hat{\mu}_{b_h} = O_p(1), \qquad \hat{\sigma}^2_{b_h} = O_p(1). \tag{60}$$

Substituting Eqs.(58)–(60) into Eq.(51) leads to the second term of Eq.(57). (Q.E.D.)

Note that the contribution of the necessary components involves the trivial redundancy (Compare the first term of Eq.(57) with Eq.(4).), because the independence between A and B prevents the VB posterior distribution from extending along the trivial degeneracy, which is a peculiarity of the LNNs.

7.2 Generalization Error

The generalization error, as well as the training error, of the VB approach can be derived in the same way as that of the SB approach, which results in a similar solution. (See Section 9.1.) [Nakajima and Watanabe, 2005b; 2006]. The existence of the singular value decomposition of any B^*A^* allows us to assume without loss of generality that A^* and B^* consist of its orthogonal row vectors and of its orthogonal column vectors, respectively. Then, we find from Lemma 1 that the KL divergence, Eq.(34), with a set of n training samples is given by

$$\begin{aligned}\tilde{G}(X^n, Y^n) &= \left\langle \frac{\|(B^*A^* - \hat{B}\hat{A})x\|^2}{2} \right\rangle_{q(x)} + O_p(n^{-3/2}) \\ &= \textstyle\sum_{h=1}^{H} \tilde{G}_h(X^n, Y^n) + O_p(n^{-3/2}), \end{aligned} \tag{61}$$

where

$$\tilde{G}_h(X^n, Y^n) = \frac{1}{2}\mathrm{tr}\Big((b^*_h a^{*t}_h - \hat{b}_h \hat{a}^t_h)^t (b^*_h a^{*t}_h - \hat{b}_h \hat{a}^t_h)\Big) \tag{62}$$

is the contribution of the h-th component. We denote by $\mathcal{W}_d(m, \Sigma, \Lambda)$ the d-dimensional Wishart distribution with m degrees of freedom, scale matrix Σ, and noncentrality matrix Λ, and abbreviate as $\mathcal{W}_d(m, \Sigma)$ the central Wishart distribution. Remember that $\theta(\cdot)$ is the indicator function, introduced in Section 3.3.

Theorem 3 *The generalization error of an LNN in the VB approach can be asymptotically expanded as*

$$\tilde{G}(n) = \tilde{\lambda} n^{-1} + O(n^{-3/2}),$$

where the generalization coefficient is given by

$$2\tilde{\lambda} = (H^*(L + l) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma'^2_h > L)\left(1 - \frac{L}{\gamma'^2_h}\right)^2 \gamma'^2_h \right\rangle_{q(\{\gamma'^2_h\})}. \tag{63}$$

Here, γ'^2_h is the h-th largest eigenvalue of a random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$, over which $\langle\cdot\rangle_{q(\{\gamma'^2_h\})}$ denotes the expectation value.

(Proof) According to Theorem 1, the difference between the VB and the ML estimators of a true component with a positive singular value is of order $O_p(n^{-1})$. Furthermore, the generalization error of the ML estimator of the component is the same as that of the regular models because of its identifiability. Hence, from Eq.(4), we obtain the first term of Eq.(63) as the contribution of the first H^* components. On the other hand, we find from Eq.(43) and Theorem 1 that, for a redundant component, identifying $RQ^{-1/2}$ with R affects the VB estimator only of order $O_p(n^{-1})$, which, hence, does not affect the generalization coefficient. We say that U is the general diagonalized matrix of an $N \times M$ matrix T if T has the following singular value decomposition: $T = \mathit{\Omega}_b U \mathit{\Omega}_a$, where $\mathit{\Omega}_a$ and $\mathit{\Omega}_b$ are an $M \times M$ and an $N \times N$ orthogonal matrices, respectively. Let D be the general diagonalized matrix of R, and D' the $(N-H^*)\times(M-H^*)$ matrix created by removing the first H^* columns and rows from D. Then, the first H^* diagonal elements of D correspond to the positive true singular value components, and D' consists only of noises. Therefore, D' is the general diagonalized matrix of $n^{-1/2}R'$, where R' is an $(N-H^*)\times(M-H^*)$ random matrix whose elements are independently subject to $\mathcal{N}_1(0,1)$, so that $R'R'^t$ is subject to $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*})$. The redundant components imitate $n^{-1/2}R'$. Hence, using Theorem 1 and Eq.(62) and noting that the distribution of the $(l-H^*)$ largest eigenvalues, which are not trivially equal to zero, of a random matrix subject to $\mathcal{W}_{L-H^*}(l-H^*, I_{L-H^*})$ are equal to that of the eigenvalues of another random matrix subject to $\mathcal{W}_{l-H^*}(L-H^*, I_{l-H^*})$, we obtain the second term of Eq.(63) as the contribution of the last $(H-H^*)$ components. Thus, we complete the proof of Theorem 3. (Q.E.D.)

Although the second term of Eq.(63) is not a simple function, it can, relatively easily, be numerically calculated by creating samples subject to the Wishart distribution. Furthermore, the more simple function approximating the term can be derived in the large scale limit when M, N, H, and H^* go to infinity in the same order, in a similar fashion to the analysis of the ML estimation [Fukumizu, 1999]. We define the following scalars:

$$\alpha = (l-H^*)/(L-H^*), \tag{64}$$

$$\beta = (H-H^*)/(l-H^*), \tag{65}$$

$$\kappa = L/(L-H^*). \tag{66}$$

Let W be a random matrix subject to $\mathcal{W}_{l-H^*}(L-H^*, I_{l-H^*})$, and $\{u_1, \ldots, u_{l-H^*}\}$ the eigenvalues of $(L-H^*)^{-1}W$. The measure of the empirical distribution of the eigenvalues is defined by

$$\delta P = (l-H^*)^{-1}\left\{\delta(u_1)+\delta(u_2)+\cdots+\delta(u_{l-H^*})\right\}, \tag{67}$$

where $\delta(u)$ denotes the Dirac measure at u. In the large scale limit, the measure, Eq.(67), converges almost everywhere to

$$p(u)du = \frac{\sqrt{(u-u_m)(u_M-u)}}{2\pi\alpha u}\theta(u_m < u < u_M)du, \tag{68}$$

where $u_m = (\sqrt{\alpha}-1)^2$ and $u_M = (\sqrt{\alpha}+1)^2$ [Watcher, 1978]. Let

$$(2\pi\alpha)^{-1}J(u_t;k) = \int_{u_t}^{\infty} u^k p(u)du \tag{69}$$

be the k-th order moment of the distribution, Eq.(68), where u_t is the lower bound of the integration range. The second term of Eq.(63) consists of the terms proportional to the minus first, the zero, and the first order moments of the eigenvalues. Because only the eigenvalues greater than L among the largest $(H - H^*)$ eigenvalues contribute to the generalization error, the moments with the lower bound $u_t = \max(\kappa, u_\beta)$ should be calculated, where u_β is the β-percentile point of $p(u)$, i.e.,

$$\beta = \int_{u_\beta}^{\infty} p(u)du = (2\pi\alpha)^{-1} J(u_\beta; 0).$$

Using the transform $s = \left(u - (u_m + u_M)/2\right)/(2\sqrt{\alpha})$, we can calculate the moments and thus obtain the following theorem:

Theorem 4 *The VB generalization coefficient of an LNN in the large scale limit is given by*

$$\begin{aligned} 2\tilde{\lambda} \sim\ & (H^*(L+l) - H^{*2}) \\ & + \frac{(L-H^*)(l-H^*)}{2\pi\alpha} \left\{ J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1) \right\}, \end{aligned} \tag{70}$$

where

$$\begin{aligned} J(s;1) &= 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s), \\ J(s;0) &= -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1} s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s + 2\alpha}{2\alpha s + \sqrt{\alpha}(1+\alpha)}, \\ J(s;-1) &= \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1} s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0 < \alpha < 1) \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1) \end{cases}, \end{aligned}$$

and $s_t = \max\left((\kappa - (1+\alpha))/2\sqrt{\alpha}, J^{-1}(2\pi\alpha\beta; 0)\right)$. *Here* $J^{-1}(\cdot; k)$ *denotes the inverse function of* $J(s;k)$.

In ordinary asymptotic analysis, one considers only situations when the amplitude of each component of the true model is zero or *distinctly-positive*. Also Theorem 3 holds only in such situations. However, as mentioned in Section 5, it is important to consider the *delicate* situations when the true map B^*A^* has tiny but non-negligible singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$. Theorem 1 still holds in such situations by replacing the second term of Eq.(54) with $o_p(n^{-1/2})$. We regard H^* as the number of *distinctly-positive* true singular values such that ${\gamma_h^*}^{-1} = o(\sqrt{n})$. Without loss of generality, we assume that B^*A^* is a non-negative, general diagonal matrix with its diagonal elements arranged in non-increasing order. Let R''^* be the true submatrix created by removing the first H^* columns and rows from B^*A^*. Then, D', defined in the proof of Theorem 3, is the general diagonalized matrix of $n^{-1/2}R''$, where R'' is a random matrix such that $R''R''^t$ is subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*}, nR''^*R''^{*t})$. Therefore, we obtain the following theorem:

Theorem 5 *The VB generalization coefficient of an LNN in the general situations when the true map B^*A^* may have delicate singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$ is given by*

$$\begin{aligned}2\tilde{\lambda} = &(H^*(L+l) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} \\ &+ \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L) \left\{ \left(1 - \frac{L}{\gamma_h''^2}\right)^2 \gamma_h''^2 - 2\left(1 - \frac{L}{\gamma_h''^2}\right) \gamma_h'' \omega_{b_h}''^t \sqrt{n} R''^* \omega_{a_h}'' \right\} \right\rangle_{q(R'')}, \end{aligned} \tag{71}$$

where γ_h'', ω_{a_h}'', and ω_{b_h}'' are the h-th largest singular value of R'', the corresponding right singular vector, and the corresponding left singular vector, respectively, of which $\langle\cdot\rangle_{q(R'')}$ denotes the expectation value over the distribution.

7.3 Training Error

Lemma 1 implies that the empirical KL divergence, Eq.(35), with a set of n training samples is given by

$$\begin{aligned}\tilde{T}(X^n, Y^n) = -\frac{1}{2}\Big\{\mathrm{tr}\Big(&(B^*A^* - \hat{B}\hat{A}_{\mathrm{MLE}})^t (B^*A^* - \hat{B}\hat{A}_{\mathrm{MLE}}) \\ &- (\hat{B}\hat{A} - \hat{B}\hat{A}_{\mathrm{MLE}})^t (\hat{B}\hat{A} - \hat{B}\hat{A}_{\mathrm{MLE}})\Big)\Big\} + O_p(n^{-3/2}). \end{aligned} \tag{72}$$

In the same way as the analysis of the generalization error, we obtain the following theorems.

Theorem 6 *The training error of an LNN in the VB approach can be asymptotically expanded as*

$$\tilde{T}(n) = \tilde{\nu} n^{-1} + O(n^{-3/2}),$$

where the training coefficient is given by

$$\begin{aligned}2\tilde{\nu} = &-(H^*(L+l) - H^{*2}) \\ &- \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L) \left(1 - \frac{L}{\gamma_h'^2}\right)\left(1 + \frac{L}{\gamma_h'^2}\right) \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \end{aligned} \tag{73}$$

Theorem 7 *The VB training coefficient of an LNN in the large scale limit is given by*

$$\begin{aligned}2\tilde{\nu} \sim &-(H^*(L+l) - H^{*2}) \\ &- \frac{(L-H^*)(l-H^*)}{2\pi\alpha} \left\{ J(s_t; 1) - \kappa^2 J(s_t; -1) \right\}. \end{aligned} \tag{74}$$

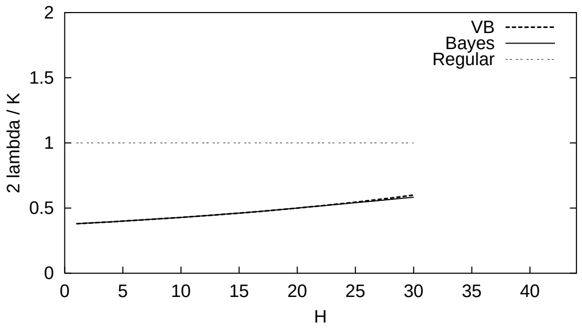

Figure 3: Free energy. ($L = 50$, $l = 30$, $H = 1, \ldots, 30$, and $H^* = 0$.) The VB and the Bayes free energies almost coincide with each other.

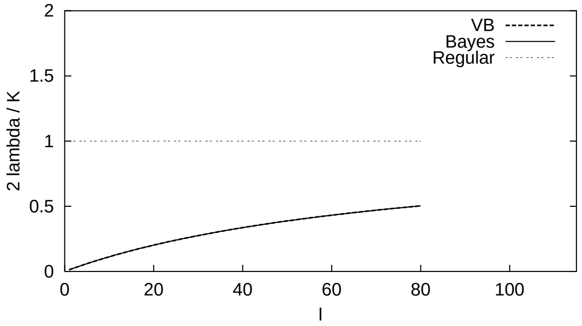

Figure 4: Free energy. ($L = 80$, $l = 1, \ldots, 80$, $H = 1$, and $H^* = 0$.) The VB and the Bayes free energies coincide with each other, as well as in Fig. 3.

Theorem 8 *The VB training coefficient of an LNN in the general situations when the true map* B^*A^* *may have delicate singular values such that* $0 < \sqrt{n}\gamma_h^* < \infty$ *is given by*

$$2\tilde{\nu} = -(H^*(L+l) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L)\left(1 - \frac{L}{\gamma_h''^2}\right)\left(1 + \frac{L}{\gamma_h''^2}\right)\gamma_h''^2 \right\rangle_{q(R'')}. \quad (75)$$

8 Theoretical Values

8.1 Free Energy

Figure 3 shows the free energy coefficients of the LNNs where $L = 50$ and $l = 30$ on the assumption that the true rank is equal to zero, $H^* = 0$. The horizontal axis indicates the rank of the learner, $H = 1, \ldots, 30$. The vertical axis indicates the coefficients normalized by the half of

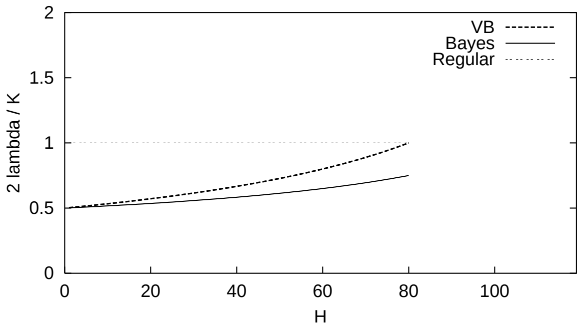

Figure 5: Free energy. ($L = l = 80$, $H = 1, \ldots, 80$, and $H^* = 0$.)

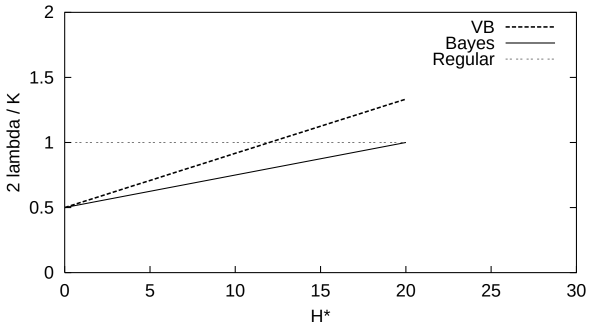

Figure 6: Free energy. ($L = 50$, $l = 30$, $H = 20$, and $H^* = 1, \ldots, 20$.)

the *essential* parameter dimension K, given by Eq.(4). The lines correspond to the free energy coefficient of the VB approach, given by Theorem 2, to that of the Bayes estimation, clarified in [Aoyagi and Watanabe, 2005], and to that of the regular models, respectively.[5] Note that the lines of the VB approach and of the Bayes estimation almost coincide with each other in Fig. 3. Figure 4, also in which the VB free energy well approximates the Bayes one, similarly shows the coefficients of LNNs where $L = 80$, $l = 1, \ldots, 80$, indicated by the horizontal axis, and $H = 1$ on the assumption that $H^* = 0$. On the other hand, the following two figures shows the cases where the VB free energy does not well approximate: Fig. 5 shows the coefficients of LNNs where $L = l = 80$, and $H = 1, \ldots, 80$, indicated by the horizontal axis, on the assumption that $H^* = 0$; and Fig. 6 shows the true rank, $H^* = 1, \ldots, 20$, dependence of the coefficients of an LNN with $L = 50$, $l = 30$, and $H = 20$ units.

Figures 7–9 show the similar cases to Fig. 4 but with $H = 10, 20, 40$, respectively. We conclude that the VB free energy behaves similarly to the Bayes one in general, and well approximates when l and L are not almost equal to each other or $H \ll l$. In Fig. 6, the VB free energy, exceptionally, strangely behaves and poorly approximates the Bayes one when H^* is large, which can, however, be regarded as a peculiarity of the LNNs, which have trivial redun-

[5] By ”regular models” we denote not LNNs but the regular models with the same parameter dimension, K, as the LNN specified by the horizontal axis.

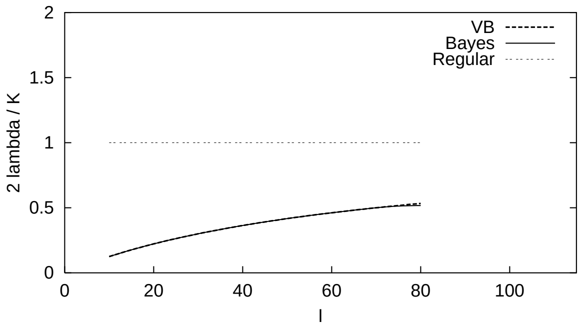

Figure 7: Free energy. ($L = 80$, $l = 1, \ldots, 80$, $H = 10$, and $H^* = 0$.)

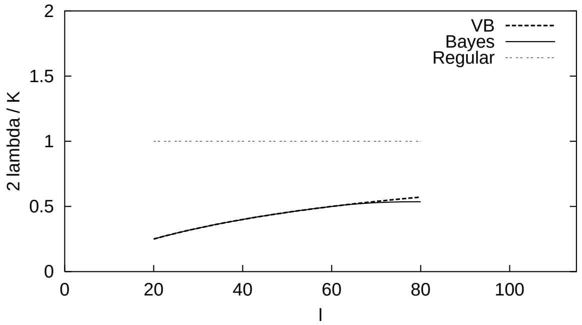

Figure 8: Free energy. ($L = 80$, $l = 1, \ldots, 80$, $H = 20$, and $H^* = 0$.)

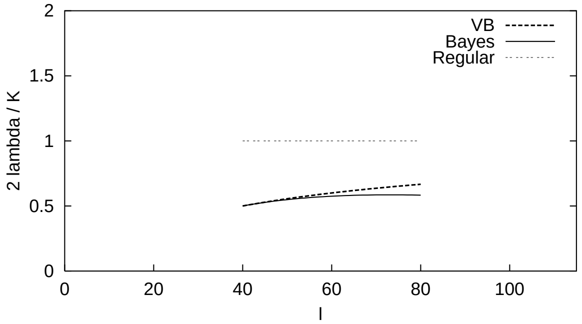

Figure 9: Free energy. ($L = 80$, $l = 1, \ldots, 80$, $H = 40$, and $H^* = 0$.)

dancy, rather than a property of the VB approach in general models. (See the last paragraph of Section 7.1.)

8.2 Generalization Error and Training Error

Figures 10–13 show the generalization and the training coefficients on the identical conditions to Figs. 3–6, respectively. The lines in the positive region correspond to the generalization coefficient of the VB approach, clarified in this paper, to that of the ML estimation, clarified in [Fukumizu, 1999], to that of the Bayes estimation, clarified in [Aoyagi and Watanabe, 2005],[6]

[6] The Bayes generalization coefficient is identical to the Bayes free energy coefficient, as mentioned in the last paragraph of Section 3.1.

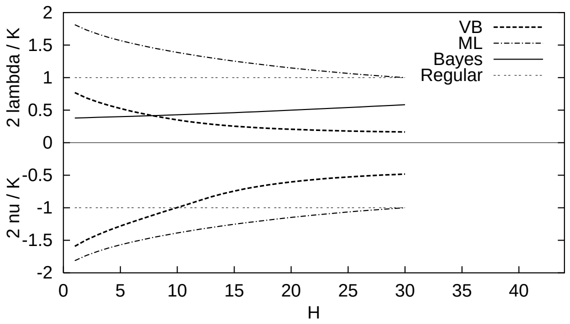

Figure 10: Generalization error (in the positive region) and training error (in the negative region). ($L = 50$, $l = 30$, $H = 1, \ldots, 30$, and $H^* = 0$.)

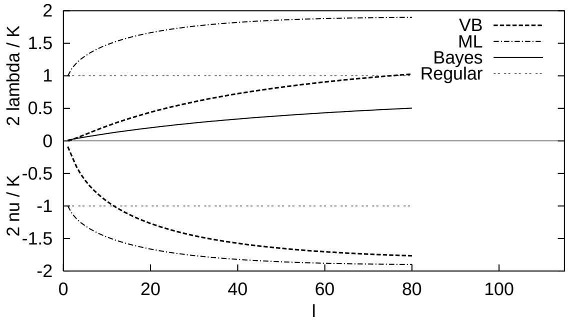

Figure 11: Generalization error and training error. ($L = 80$, $l = 1, \ldots, 80$, $H = 1$, and $H^* = 0$.)

and to that of the regular models, respectively; while the lines in the negative region correspond to the training coefficient of the VB approach, to that of the ML estimation, and to that of the regular models, respectively.[7] Unfortunately the Bayes training error has not been clarified yet. The results in Figs. 10–13 have been calculated in the large scale approximation, i.e., by using Theorems 4 and 7. We have also numerically calculated them by using Theorems 3 and 6, and thus found that the both results almost coincide with each other so that we can hardly distinguish.

We see in Figs. 10–13 that the VB approach provides good generalization performance comparable to, or in some cases better than, the Bayes estimation. We also see that the VB generalization coefficient significantly differs from the Bayes one even in the cases where the VB free energy well approximates the Bayes one. (Compare, for example, Figs. 3 and 4 with the positive regions of Figs. 10 and 11.) Furthermore, we see in Figs. 5 and 12 that, when $H \ll l$, the VB approach provides much worse generalization performance than the Bayes estimation, while the VB free energy well approximates the Bayes one; and that, when $H \sim l$, the VB approach provides much better generalization performance, while the VB free energy is signifi-

[7] By the term ”of the regular models” we mean both ”of the ML estimation and of the Bayes estimation in the regular models.”

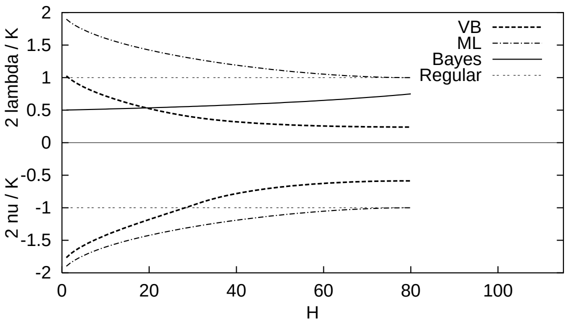

Figure 12: Generalization error and training error. ($L = l = 80$, $H = 1, \ldots, 80$, and $H^* = 0$.)

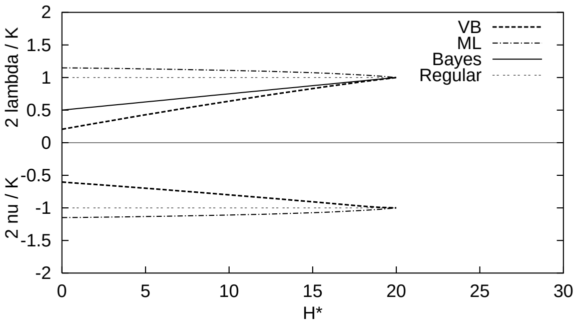

Figure 13: Generalization error and training error. ($L = 50$, $l = 30$, $H = 20$, and $H^* = 1, \ldots, 20$.)

cantly larger than the Bayes one. This can throw a doubt on model selection by minimizing the VB free energy to obtain better generalization performance. (See Section 9.2.)

We see in Figs. 10 and 12 that the VB generalization coefficient depends on H similarly to the ML one. Moreover, we see in Fig. 12 that, when $H = 1$, the VB generalization coefficient exceeds that of the regular models, which the Bayes one never exceeds [Watanabe, 2001b]. That is for the following reason: because of the asymptotic equivalence to the shrinkage estimation, the VB approach would approximate the ML estimation and hence provide poor generalization performance in models where the ML estimator of a redundant component would go out of the effective range of shrinkage, i.e., be much greater than $\sqrt{L/n}$. When $(l - H^*) \gg (H - H^*)$, the $(H - H^*)$ largest eigenvalues of a random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$ are much greater than L, because of selection from a large number of random variables subject to the non-compact support distribution. Therefore, the eigenvalues $\{\gamma_h'^2\}$ in Theorem 3 go out of the effective range of shrinkage. It can be said that the VB approach has not only a property similar to the Bayes estimation, i.e., suppression of overfitting by the JS type shrinkage, but also another property similar to the ML estimation, i.e., acceleration of overfitting by selection of the largest singular values of a random matrix.

Figure 13 shows that, in this LNN, the VB approach provides no worse generalization performance than the Bayes estimation regardless of H^*. This, however, does not mean the dom-

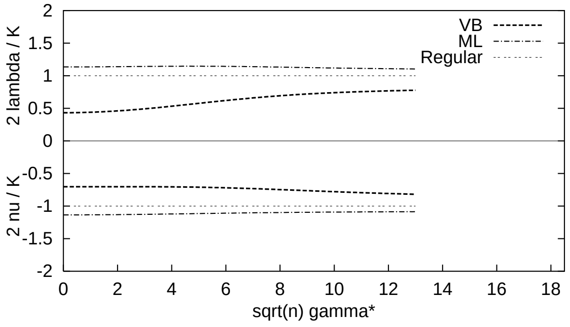

Figure 14: *Delicate* situations.

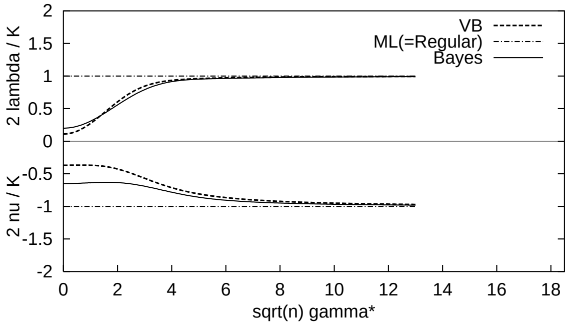

Figure 15: *Delicate* situations in a single-output LNN.

ination of the VB approach over the Bayes estimation. The consideration of *delicate* situations in the following denies the domination. Using Theorems 5 and 8, we can numerically calculate the VB, as well as the ML,[8] generalization and training coefficients in the *delicate* situations when the true distribution is near the singularities. Figure 14 shows the coefficients of an LNN where $L = 50$, $l = 30$, and $H = 20$ on the assumption that the true map consists of $H^* = 5$ *distinctly-positive* component, the ten *delicate* components whose singular values are identical to each other, and the other five null components. The horizontal axis indicates $\sqrt{n}\gamma^*$, where $\gamma_h^* = \gamma^*$ for $h = 6, \ldots, 15$. The Bayes generalization error in the *delicate* situations was clarified in [Watanabe and Amari, 2003], but unfortunately, only in single-output (SO) LNNs, i.e., $N = H = 1$.[9] Figure 15 shows the coefficients of an SOLNN with $M = 5$ input units on the assumption that $H^* = 0$ and the one true singular value, indicated by the horizontal axis, is *delicate*. We see that, in some *delicate* situations, the VB approach provides worse generalization performance than the Bayes estimation, although this is one of the cases that the VB approach seemed to dominate the Bayes estimation, without consideration of *delicate* situations. Note

[8] All of the theorems in this paper except Theorem 2 can be modified for the ML estimation by letting $L = 0$.

[9] An SOLNN is regarded as a regular model from the viewpoint of the ML estimation because the transform $b_1 a_1 \mapsto w$, where $w \in \mathbb{R}^M$, makes the model identifiable, and hence the ML generalization coefficient is identical to that of the regular models, as shown in Fig. 15. Nevertheless, Fig. 15 shows that an SOLNN has a property of unidentifiable models from the viewpoint of the Bayesian learning methods.

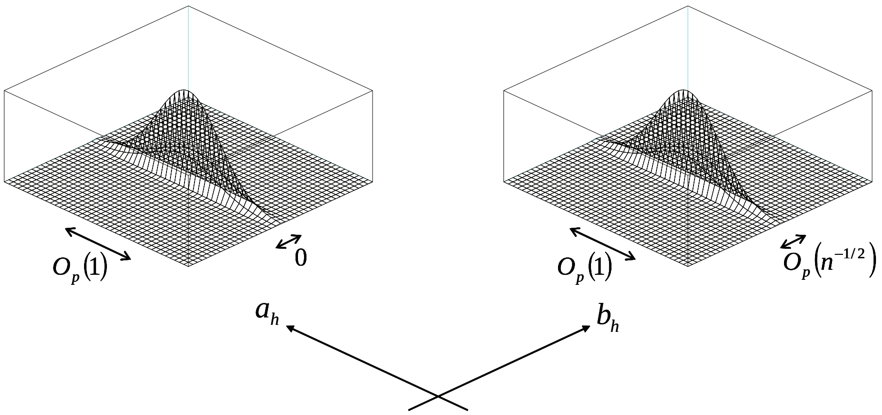

Figure 16: SB (MIP) posterior (left-hand side) and VB posterior (right-hand side) for $H^* < h \leq H$ when $M > N$.

that the ordinary asymptotic cases when $H^* = 5$ and when $H^* = 15$ (in Fig. 13) correspond to the case that $\sqrt{n}\gamma^* = 0$ and the case that $\sqrt{n}\gamma^* \to \infty$ in Fig. 14, respectively. It is expected that, also in Fig.14, there must be the region where the VB approach provides greater generalization error than the Bayes estimation, because the Bayes estimation is admissible. (See Section 9.3.) In other words, Fig. 13 shows the generalization error variation when the number, H^*, of positive true components increases one by one, and, in that case, the VB approach never provides worse generalization performance for any H^* than the Bayes estimation; while Fig. 14 shows the generalization error variation when the amplitudes of ten true components simultaneously increase, and, in that case, the VB approach is expected to provide worse generalization performance in some region.

9 Discussion

9.1 Relation to Subspace Bayes Approach

Interestingly, the solution of the VB approach is similar to that of the subspace Bayes (SB) approach [Nakajima and Watanabe, 2005b; 2006]. The VB solution is, actually, asymptotically equivalent to that of the MIP version of SB approach, the empirical Bayes (EB) approach where the output parameter matrix is regarded as a hyperparameter, when $M \geq N$, and to that of the MOP version, the EB approach where the input parameter matrix is regarded as a hyperparameter, when $M \leq N$. The form of the posterior distribution of the parameters corresponding to the redundant components implies the reason of the equivalence. The SB posterior distribution extends with its variance of order $O_p(1)$ in the space of the parameters, i.e., the ones not regarded as hyperparameters; while the VB posterior distribution extends with its variance of order $O_p(1)$ in the larger dimension parameter space, either the input one or the output one, as we can find in the proof of Theorem 2. So, the similarity between the solutions is natural, and it can be said that the VB approach automatically selects the larger dimension space as the marginalized space, and that the dimension of the space in which the singularities allow the posterior distribution to extend is essentially related to the degree of shrinkage. Figure 16 illustrates the posterior distribution of the MIP version of the SB approach (left-hand side), and that of the VB approach (right-hand side) of a redundant component when $M > N$.[10]

The relation between the EB approach and the JS estimation was discussed in [Efron and Morris, 1973], as mentioned in Section 1. In this paper, we have discovered the relation among the VB approach, the EB approach, and the JS type shrinkage estimation.

9.2 Model Selection in VB Approach

In the framework of the Bayes estimation, the marginal likelihood, Eq.(7), can be regarded as the likelihood of the *ensemble* of the models. Following the concept of *likelihood* in statistics, we can accept that the *ensemble* of models with the maximum marginal likelihood is the most likely. Thus, the model selection method minimizing the free energy, Eq.(10), was proposed [Efron and Morris, 1973; Schwarz, 1978; Akaike, 1980; MacKay, 1992]. In the VB approach, the method where the model with the minimum VB free energy is selected was proposed [Attias, 1999].

Although it is known that the model with the minimum free energy does not necessarily provide the minimum generalization error even in the Bayes estimation and even in regular models, we usually expect that the selected model would also provide a small generalization error. However, we have shown in this paper that the free energy and the generalization error are less simply related with each other in the VB approach, unlike in the Bayes estimation. This can imply that the increase of the free energy caused by overfitting with redundant components less accurately indicates that of the generalization error in the VB approach than in the Bayes estimation, and throw a doubt on this model selection method.

9.3 Consistency with Superiority of Bayes Estimation

The Bayes estimation is said to be superior to any other learning method in a certain meaning. Consider the case that we apply a learning method to many applications, where the true parameter w^* is different in each application and subject to $q(w^*)$. Note that the true distribution in each application is written as $q(y|x) = p(y|x, w^*)$. Although we have been abbreviating it except in Section 3.3, the generalization error, Eq.(13), naturally depends on the true distribution, so we here denote the dependence explicitly as follows: $G(n; w^*)$. Let

$$\bar{G}(n) = \langle G(n; w^*) \rangle_{q(w^*)} \tag{76}$$

be the average generalization error over $q(w^*)$.

[10] Note that the SB posterior distribution is the delta function with respect to b_h, so that it has infinite values in reality.

Proposition 1 *When we know and use the true prior distribution, $q(w^*)$, the Bayes estimation minimizes the average generalization error over $q(w^*)$, i.e.,*

$$\bar{G}(n) \leq \bar{G}_{\mathit{Other}}(n), \tag{77}$$

where $\bar{G}_{\mathit{Other}}(n)$ denotes the average generalization error of an arbitrary learning method.

However, we consider the situations where we need model selection, i.e., we do not know even the true dimension of the parameter space. Accordingly, it can happen that an approximation method of the Bayes estimation provides better generalization performance than the Bayes estimation, as shown in this paper. On the other hand, Proposition 1 naturally leads to the admissibility of the Bayes estimation, i.e., there is no learning method dominating the Bayes estimation. The analysis of the *delicate* situation has consistently denied the domination of the VB approach over the Bayes estimation in the last paragraph of Section 8.2.

9.4 Features of VB Approach

To summarize, we can say that the VB approach has the two aspects caused by the *ML type effect* and the *Bayesian effect* of the singularities, itemized in Section 4. In LNNs, the *ML type effect* is observed as the acceleration of overfitting by selection of the largest singular values of a random matrix; while the *Bayesian effect* is observed as the JS type *shrinkage*. We have also shown in Figs. 11 and 12 that the one of the most preferable properties of the Bayes estimation, i.e., $2\lambda \leq K$, does not hold in the VB approach, i.e., $2\tilde{\lambda}$ can exceeds the parameter dimension, K, hence the generalization coefficient of the regular models. We conjecture that the two aspects are essentially caused by the localization of the posterior distribution, which we can see in Fig. 1. Because of the independence between the parameters of different layers, the posterior distribution cannot extend so as to fill up the chink of singularities, which changes the strength of the *Bayesian effect*, sometimes increases, and sometimes decreases.

The restriction on the VB posterior distribution affects the generalization performance. However, the variation of the restriction is limited if we like to obtain a simple iterative algorithm. Especially, the independence between the parameters of different layers is indispensable to simplify the calculation. Actually, the restriction applied in models with hidden variables, such as mixture models, hidden Markov models, etc., is very similar to the one applied in LNNs in this paper. In [Attias, 1999], the VB posterior distribution is restricted such that the parameters and the hidden variables are independent of each other. That restriction results in the independence between the parameters of different layers, and consequently, leads to the localization of the posterior distribution. In fact, the order of the extent of the VB posterior distribution in mixture models when we use the prior distribution having positive values on the singularities has recently been derived to be similar to that in LNNs, [Watanabe and Watanabe, 2006]. So, we expect that the two aspects of the VB approach would hold also in more general unidentifiable models. In addition to that, in models with hidden variables, the other restriction that we applied in LNNs, i.e., the independence between the parameters of different components, is also guaranteed when we introduce the hidden variables and substitute the likelihood of the *complete* data for the original *marginal* likelihood, in the same way as the derivation of the EM algorithm [Dempster *et al.*, 1977; Attias, 1999].

Theorem 1 says that increasing the smaller dimension, l, does not increase the degree of shrinkage unless $L = l$, but Theorem 3 says that it increases the number of the random variables, i.e., the number of the eigenvalues of a random matrix subject to $\mathcal{W}_{l-H^*}(L - H^*, I_{l-H^*})$, among which the redundant components of the VB estimator choose. Therefore, we can say that the VB generalization error is very small when the difference between the parameter space dimensions of the different layers is very large, i.e., $l/L \ll 1$. So, we conjecture that the advantage of the VB approach over the EM algorithm would be enhanced in mixture models, where the dimension of the upper layer parameters, i.e., the number of the mixing coefficients, is one per component and that of the lower layer parameters, i.e., the number of the parameters that each component has, is usually more.

However, in general, non-linearity increases the *ML type effect* of singularities. In (general) neural networks, Eq.(1), for example, we expect that the non-linearity of the activation function, $\psi(\cdot)$, would extend the range of basis selection and hence increase the generalization error. Because an LNN can be embedded in a finite dimensional regular model, the *ML type effect* of the singularities is relatively soft, and its ML estimator is in a finite region. On the other hand, the ML estimator can diverge in some unidentifiable models. An important question to be solved is how the VB approach behaves in the models where the ML estimator diverges because of the *ML type effect*, that is, whether the *Bayesian shrinkage effect* leashes the VB estimator in a finite region or not.

As mentioned in Section 1, the VB free energies in some unidentifiable models have been clarified [Watanabe and Watanabe, 2006; Hosino *et al.*, 2005; Nakano and Watanabe, 2005] and shown to behave similarly to the VB free energy in LNNs, except the peculiarity of LNNs caused by the trivial redundancy, i.e., the strange H^* dependence discussed in the last paragraph of Section 8.1. The fact shown in this paper that, in LNNs, the VB free energy and the VB generalization error are less simply related to each other implies that, also in other models, we cannot expect the VB approach to provide the similar generalization performance to the Bayes estimation, even if it provides the similar free energy.

10 Conclusions and Future Work

We have proved that, in three-layer linear neural networks, the variational Bayes (VB) approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, and theoretically clarified its free energy, generalization error, and training error. We have thus concluded that the VB approach has not only a property similar to the Bayes estimation but also another property similar to the maximum likelihood estimation. We have also shown that, unlike in the Bayes estimation, the free energy and the generalization error are less simply related to each other, and discussed the relation between the VB approach and the subspace Bayes (SB) approach.

Analysis of the VB generalization performance of other unidentifiable models is future work. Consideration of the relation between the VB approach and the SB approach in other models is also future work. We think that any kind of similarity between the approaches may hold not only in LNNs.

Acknoledgement

The authors would like to thank anonymous reviewers for the meaningful comments that polished this paper. The authors would also like to thank Kazuo Ushida, Masahiro Nei, and Nobutaka Magome of Nikon Corporation for encouragement to research this subject.

A Proof of Theorem 1

Both tr(Σ_{a_h}) and tr(Σ_{b_h}) are of no greater order than 1 because of the prior distributions, Eqs.(39) and (40). Both of them are also not equal to zero, otherwise the generalized free energy, Eq.(51), diverges to infinity with finite n. Consequently, the generalized free energy diverges to infinity when either $\|\mu_{a_h}\|$ or $\|\mu_{b_h}\|$ goes to infinity. Hence, the optimum values of μ_{a_h}, μ_{b_h}, Σ_{a_h}, and Σ_{b_h} necessarily satisfy the variational condition, Eqs.(47)–(50). Combining Eqs.(47) and (49), we have

$$\hat{\mu}_{a_h} = n^2 \hat{\Sigma}_{a_h} R^t \hat{\Sigma}_{b_h} R \hat{\mu}_{a_h}, \tag{78}$$

$$\hat{\mu}_{b_h} = n^2 \hat{\Sigma}_{b_h} R \hat{\Sigma}_{a_h} R^t \hat{\mu}_{b_h}, \tag{79}$$

and thus find that $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are an eigenvector of $R^t R$ and an eigenvector of $R\hat{\Sigma}_{a_h}R^t$, respectively, or $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$. Hereafter, separately considering the necessary components, which imitate the true ones with positive singular values, and the redundant components, we will find the solution of the variational condition that minimizes the generalized free energy, Eq.(51).

For a necessary component, $h \leq H^*$, the (observed) singular value of R is of order $O_p(1)$. Hence, the free energy, Eq.(51), can be minimized only when both $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are of order $O_p(1)$. Therefore, the variational condition, Eqs.(47)–(50), is approximated as follows:

$$\hat{\mu}_{a_h} = \|\hat{\mu}_{b_h}\|^{-2} Q^{-1} R^t \hat{\mu}_{b_h} + O_p(n^{-1}), \tag{80}$$

$$\hat{\Sigma}_{a_h} = n^{-1} \|\hat{\mu}_{b_h}\|^{-2} Q^{-1} + O_p(n^{-2}), \tag{81}$$

$$\hat{\mu}_{b_h} = (\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} R \hat{\mu}_{a_h} + O_p(n^{-1}), \tag{82}$$

$$\hat{\Sigma}_{b_h} = n^{-1} (\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} I_N + O_p(n^{-2}). \tag{83}$$

Thus, we obtain the VB estimator of the necessary component:

$$\hat{\mu}_{b_h} \hat{\mu}_{a_h}^t = \omega_{b_h} \omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \tag{84}$$

On the other hand, for a redundant component, $h > H^*$, Eqs.(43), (52), and (53) allow us to approximate the variational condition, Eqs.(47)–(50), as follows:

$$\hat{\mu}_{a_h} = n\hat{\sigma}^2_{a_h} R^t \hat{\mu}_{b_h}(1 + O_p(n^{-1/2})), \tag{85}$$

$$\hat{\sigma}^2_{a_h} = (n(\|\hat{\mu}_{b_h}\|^2 + N\hat{\sigma}^2_{b_h}) + c_a^{-2})^{-1}, \tag{86}$$

$$\hat{\mu}_{b_h} = n\hat{\sigma}^2_{b_h} R \hat{\mu}_{a_h}(1 + O_p(n^{-1/2})), \tag{87}$$

$$\hat{\sigma}^2_{b_h} = (n(\|\hat{\mu}_{a_h}\|^2 + M\hat{\sigma}^2_{a_h}) + c_b^{-2})^{-1}. \tag{88}$$

Combining Eqs.(86) and (88), we have

$$\hat{\sigma}^2_{a_h} = \frac{-(n\hat{\eta}_h^2 - (M-N)) + \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN}}{2nM(\|\hat{\mu}_{b_h}\|^2 + n^{-1}c_a^{-2})}, \tag{89}$$

$$\hat{\sigma}^2_{b_h} = \frac{-(n\hat{\eta}_h^2 + (M-N)) + \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN}}{2nN(\|\hat{\mu}_{a_h}\|^2 + n^{-1}c_b^{-2})}, \tag{90}$$

where

$$\hat{\eta}_h^2 = \left(\|\hat{\mu}_{a_h}\|^2 + \frac{c_b^{-2}}{n}\right)\left(\|\hat{\mu}_{b_h}\|^2 + \frac{c_a^{-2}}{n}\right). \tag{91}$$

We consider two possibilities of solutions:

1. Such that $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$:

   Obviously, the following equations satisfy Eqs.(85) and (87):

$$\hat{\mu}_{a_h} = 0, \tag{92}$$

$$\hat{\mu}_{b_h} = 0. \tag{93}$$

   Substituting Eqs.(92) and (93) into Eqs.(89) and (90), we easily find the following solution:

   (a) When $M > N$:

$$\hat{\sigma}^2_{a_h} = \frac{M - N}{Mc_a^{-2}} + O_p(n^{-1}), \tag{94}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_a^{-2}}{n(M - N)} + O_p(n^{-2}). \tag{95}$$

   (b) When $M = N$:

$$\hat{\sigma}^2_{a_h} = \frac{c_a}{c_b\sqrt{nM}} + O_p(n^{-1}), \tag{96}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_b}{c_a\sqrt{nM}} + O_p(n^{-1}). \tag{97}$$

(c) When $M < N$:

$$\hat{\sigma}_{a_h}^2 = \frac{c_b^{-2}}{n(N-M)} + O_p(n^{-2}), \tag{98}$$

$$\hat{\sigma}_{b_h}^2 = \frac{N-M}{Nc_b^{-2}} + O_p(n^{-1}). \tag{99}$$

2. Such that $\|\hat{\mu}_{a_h}\|, \|\hat{\mu}_{b_h}\| > 0$:

Combining Eqs.(85) and (87), we find that $\hat{\mu}_{a_h}$ and $\hat{\mu}_{b_h}$ are the right and the corresponding left singular vectors of R, respectively. Let h-th largest singular value component of R correspond to the h-th component of the estimator. Then we have

$$\hat{\mu}_{a_h} = \|\hat{\mu}_{a_h}\|\omega_{a_h}, \tag{100}$$

$$\hat{\mu}_{b_h} = \|\hat{\mu}_{b_h}\|\omega_{b_h}. \tag{101}$$

Substituting Eqs.(89), (90), (100), and (101) into Eqs.(85) and (87), we have

$$\frac{2MN}{n\gamma_h^2} = n\hat{\eta}_h^2 + M + N - \sqrt{(n\hat{\eta}_h^2 + M + N)^2 - 4MN} + O_p(n^{-1/2}), \tag{102}$$

$$(Mc_a^{-2}\hat{\delta}_h - Nc_b^{-2}\hat{\delta}_h^{-1})(1 + O_p(n^{-1/2})) = n(M-N)(\gamma_h - \hat{\gamma}_h), \tag{103}$$

where

$$\hat{\gamma}_h = \|\hat{\mu}_{a_h}\|\|\hat{\mu}_{b_h}\|, \tag{104}$$

$$\hat{\delta}_h = \frac{\|\hat{\mu}_{a_h}\|}{\|\hat{\mu}_{b_h}\|}. \tag{105}$$

Equation (102) implies that $\xi = (n\gamma_h^2)^{-1}$ is a solution of the following equation:

$$MN\xi^2 - (n\hat{\eta}_h^2 + M + N)\xi + 1 = O_p(n^{-1/2}). \tag{106}$$

Thus, we find that Eq.(102) has the solution below when and only when $n\gamma_h^2 \geq L$:

$$\hat{\eta}_h^2 = \left(1 - \frac{M}{n\gamma_h^2}\right)\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h^2 + O_p(n^{-3/2}). \tag{107}$$

By using Eqs.(104) and (105), the definition of $\hat{\eta}_h$, Eq.(91), can be written as follows:

$$\hat{\eta}_h^2 = \left(1 + \frac{c_b^{-2}}{n\hat{\gamma}_h\hat{\delta}_h}\right)\left(1 + \frac{c_a^{-2}}{n\hat{\gamma}_h\hat{\delta}_h^{-1}}\right)\hat{\gamma}_h^2, \tag{108}$$

Hereafter, we will find the simultaneous solution of Eqs.(103), (107), and (108) in order to obtain the solution that exists only when $n\gamma_h^2 \geq L$. Equations (107) and (108) imply that $\gamma_h^2 > \hat{\eta}_h^2 > \hat{\gamma}_h^2$ and that $(\gamma_h^2 - \hat{\gamma}_h^2)$ is of order $O_p(n^{-1})$. We then find that $(\gamma_h - \hat{\gamma}_h)$ is of order $O_p(n^{-1/2})$ because γ_h is of order $O_p(n^{-1/2})$.

(a) When $M > N$:

In this case, Eq.(103) can be approximated as follows:

$$\hat{\delta}_h = \frac{n(M-N)(\gamma_h - \hat{\gamma}_h)}{Mc_a^{-2}} + O_p(n^{-1/2}). \tag{109}$$

Substituting Eqs.(107) and (109) into Eq.(108), we have

$$\hat{\gamma}_h^2 + \frac{(M-N)(\gamma_h - \hat{\gamma}_h)}{M}\hat{\gamma}_h - \left(1 - \frac{M}{n\gamma_h^2}\right)\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h^2 = O_p(n^{-2}), \tag{110}$$

and hence

$$\left(\hat{\gamma}_h - \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)\left(\hat{\gamma}_h + \frac{M}{N}\left(1 - \frac{N}{n\gamma_h^2}\right)\gamma_h\right) = O_p(n^{-2}). \tag{111}$$

Therefore, we obtain the following solution with respect to $\hat{\gamma}_h$ and $\hat{\delta}_h$:

$$\hat{\gamma}_h = \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h + O_p(n^{-3/2}), \tag{112}$$

$$\hat{\delta}_h = \frac{(M-N)}{c_a^{-2}\gamma_h} + O_p(n^{-1/2}). \tag{113}$$

We thus obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{M}{n\gamma_h^2}\right)\frac{M-N}{c_a^{-2}}\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \tag{114}$$

$$\hat{\sigma}_{a_h}^2 = \frac{M-N}{c_a^{-2}n\gamma_h^2} + O_p(n^{-1}), \tag{115}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{M}{n\gamma_h^2}\right)\frac{c_a^{-2}}{M-N}\right)^{1/2}\gamma_h\omega_{b_h} + O_p(n^{-1}), \tag{116}$$

$$\hat{\sigma}_{b_h}^2 = \frac{c_a^{-2}}{n(M-N)} + O_p(n^{-2}). \tag{117}$$

(b) When $M = N$:

In this case, we find from Eq.(103) that

$$\hat{\delta}_h = \frac{c_a}{c_b}. \tag{118}$$

Substituting Eq.(118) and then Eq.(107) into Eq.(108), we have

$$\begin{aligned}\hat{\gamma}_h &= \hat{\eta}_h + O_p(n^{-1}) \\ &= \left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h + O_p(n^{-1}).\end{aligned} \tag{119}$$

We thus obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\frac{c_a}{c_b}\left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \tag{120}$$

$$\hat{\sigma}_{a_h}^2 = \frac{c_a}{c_b n \gamma_h} + O_p(n^{-1}), \tag{121}$$

$$\hat{\mu}_{b_h} = \left(\frac{c_b}{c_a}\left(1 - \frac{M}{n\gamma_h^2}\right)\gamma_h\right)^{1/2}\omega_{b_h} + O_p(n^{-1}), \tag{122}$$

$$\hat{\sigma}_{b_h}^2 = \frac{c_b}{c_a n \gamma_h} + O_p(n^{-1}). \tag{123}$$

(c) When $M < N$:

In exactly the same way as the case when $M > N$, we obtain the following solution:

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{N}{n\gamma_h^2}\right)\frac{c_b^{-2}}{N - M}\right)^{1/2}\gamma_h\omega_{a_h} + O_p(n^{-1}), \tag{124}$$

$$\hat{\sigma}_{a_h}^2 = \frac{c_b^{-2}}{n(N - M)} + O_p(n^{-2}), \tag{125}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{N}{n\gamma_h^2}\right)\frac{N - M}{c_b^{-2}}\right)^{1/2}\omega_{b_h} + O_p(n^{-1}), \tag{126}$$

$$\hat{\sigma}_{b_h}^2 = \frac{N - M}{c_b^{-2} n \gamma_h^2} + O_p(n^{-1}). \tag{127}$$

We can find that, when the solution such that $\|\hat{\mu}_{a_h}\|, \|\hat{\mu}_{b_h}\| > 0$, Eqs.(114)–(117) and (120)–(127), exists, it makes the VB free energy, Eq.(51), smaller than the solution such that $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$, Eqs.(92)–(99). Hence, we arrive at the following solution:

1. When $M > N$,

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{L}{L_h'}\right)\frac{L - l}{c_a^{-2}}\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \tag{128}$$

$$\hat{\sigma}_{a_h}^2 = \frac{L - l}{c_a^{-2} L_h'} + O_p(n^{-1}), \tag{129}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{L}{L_h'}\right)\frac{c_a^{-2}}{L - l}\right)^{1/2}\gamma_h\omega_{b_h} + O_p(n^{-1}), \tag{130}$$

$$\hat{\sigma}_{b_h}^2 = \frac{c_a^{-2}}{n(L - l)} + O_p(n^{-2}). \tag{131}$$

2. When $M = N$,

$$\hat{\mu}_{a_h} = \left(\frac{c_a}{c_b} \left(1 - \frac{L}{L'_h} \right) \gamma_h \right)^{1/2} \omega_{a_h} + O_p(n^{-1}), \tag{132}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_a}{c_b \sqrt{n L'_h}} + O_p(n^{-1}), \tag{133}$$

$$\hat{\mu}_{b_h} = \left(\frac{c_b}{c_a} \left(1 - \frac{L}{L'_h} \right) \gamma_h \right)^{1/2} \omega_{b_h} + O_p(n^{-1}), \tag{134}$$

$$\hat{\sigma}^2_{b_h} = \frac{c_b}{c_a \sqrt{n L'_h}} + O_p(n^{-1}). \tag{135}$$

3. When $M < N$,

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{c_b^{-2}}{L - l} \right)^{1/2} \gamma_h \omega_{a_h} + O_p(n^{-1}), \tag{136}$$

$$\hat{\sigma}^2_{a_h} = \frac{c_b^{-2}}{n(L - l)} + O_p(n^{-2}), \tag{137}$$

$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{L}{L'_h} \right) \frac{L - l}{c_b^{-2}} \right)^{1/2} \omega_{b_h} + O_p(n^{-1}), \tag{138}$$

$$\hat{\sigma}^2_{b_h} = \frac{L - l}{c_b^{-2} L'_h} + O_p(n^{-1}). \tag{139}$$

Selecting the largest singular value components minimizes the free energy, Eq.(51). Hence, combining Eq.(84) with the fact that $L L'^{-1}_h = O_p(n^{-1})$ for the necessary components, and the solution above with Eq.(43), we obtain the VB estimator in Theorem 1. (Q.E.D.)

References

[Akaike, 1974] H. Akaike. A New Look at Statistical Model. *IEEE Trans. on Automatic Control*, 19(6):716–723, 1974.

[Akaike, 1980] H. Akaike. Likelihood and Bayes Procedure. In J. M. Bernald, editor, *Bayesian Statistics*, pages 143–166. University Press, 1980.

[Amari *et al.*, 2006] S. Amari, H. Park, and T. Ozeki. Singularities Affect Dynamics of Learning in Neuromanifolds. *Neural Computation*, 18:1007–1065, 2006.

[Aoyagi and Watanabe, 2005] M. Aoyagi and S. Watanabe. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. *Neural Networks*, 18(7):924–933, 2005.

[Attias, 1999] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proc. of UAI*, 1999.

[Baldi and Hornik, 1995] P. F. Baldi and K. Hornik. Learning in Linear Neural Networks: a Survey. *IEEE Trans. on Neural Networks*, 6(4):837–858, 1995.

[Bickel and Chernoff, 1993] P. Bickel and H. Chernoff. Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem. pages 83–96. Wiley Eastern Limited, 1993.

[Cramer, 1949] H. Cramer. *Mathematical Methods of Statistics*. University Press, Princeton, 1949.

[Dacunha-Castelle and Gassiat, 1997] D. Dacunha-Castelle and E. Gassiat. Testing in Locally Conic Models, and Application to Mixture Models. *Probability and Statistics*, 1:285–317, 1997.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood for Incomplete Data Via the EM Algorithm. *J. R. Statistical Society*, 39-B:1–38, 1977.

[Efron and Morris, 1973] B. Efron and C. Morris. Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach. *J. of Am. Stat. Assoc.*, 68:117–130, 1973.

[Fukumizu, 1999] K. Fukumizu. Generalization Error of Linear Neural Networks in Unidentifiable Cases. In *Proc. of ALT*, pages 51–62. Springer, 1999.

[Fukumizu, 2003] K. Fukumizu. Likelihood Ratio of Unidentifiable Models and Multilayer Neural Networks. *Annals of Statistics*, 31(3):833–851, 2003.

[Ghahramani and Beal, 2001] Z. Ghahramani and M. J. Beal. Graphical Models and Variational Methods. In *Advanced Mean Field Methods*, pages 161–177. MIT Press, 2001.

[Hagiwara, 2002] K. Hagiwara. On the Problem in Model Selection of Neural Network Regression in Overrealizable Scenario. *Neural Computation*, 14:1979–2002, 2002.

[Hartigan, 1985] J. A. Hartigan. A Failure of Likelihood Ratio Asymptotics for Normal Mixtures. In *Proc. of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*, pages 807–810, 1985.

[Hinton and van Camp, 1993] G. E. Hinton and D. van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proc. of COLT*, pages 5–13, 1993.

[Hosino *et al.*, 2005] T. Hosino, K. Watanabe, and S. Watanabe. Stochastic Complexity of Variational Bayesian Hidden Markov Models. In *Proc. of IJCNN*, 2005.

[Jaakkola and Jordan, 2000] T. S. Jaakkola and M. I. Jordan. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, 10:25–37, 2000.

[James and Stein, 1961] W. James and C. Stein. Estimation with Quadratic Loss. In *Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob.*, pages 361–379, 1961.

[Kuriki and Takemura, 2001] S. Kuriki and A. Takemura. Tail Probabilities of the Maxima of Multilinear Forms and Their Applications. *Annals of Statistics*, 29(2):328–371, 2001.

[Levin *et al.*, 1990] E. Levin, N. Tishby, and S. A. Solla. A Statistical Approaches to Learning and Generalization in Layered Neural Networks. In *Proc. of IEEE*, volume 78, pages 1568–1674, 1990.

[MacKay, 1992] D. J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(2):415–447, 1992.

[MacKay, 1995] D. J. C. MacKay. Developments in Probabilistic Modeling with Neural Networks—Ensemble Learning. In *Proc. of the 3rd Ann. Symp. on Neural Networks*, pages 191–198, 1995.

[Nakajima and Watanabe, 2005a] S. Nakajima and S. Watanabe. Generalization Error and Free Energy of Variational Bayes Approach of Linear Neural Networks. In *Proc. of ICONIP*, pages 55–60, Taipei, Taiwan, 2005.

[Nakajima and Watanabe, 2005b] S. Nakajima and S. Watanabe. Generalization Error of Linear Neural Networks in an Empirical Bayes Approach. In *Proc. of IJCAI*, pages 804–810, Edinburgh, U.K., 2005.

[Nakajima and Watanabe, 2006] S. Nakajima and S. Watanabe. Generalization Performance of Subspace Bayes Approach in Linear Neural Networks. *IEICE Trans.*, E89-D(3):1128–1138, 2006.

[Nakano and Watanabe, 2005] N. Nakano and S. Watanabe. Stochastic Complexity of Layered Neural Networks in Mean Field Approximation. In *Proc. of ICONIP*, Taipei, Taiwan, 2005.

[Reinsel and Velu, 1998] G. C. Reinsel and R. P. Velu. *Multivariate Reduced-Rank Regression*. Springer, 1998.

[Rissanen, 1986] J. Rissanen. Stochastic Complexity and Modeling. *Annals of Statistics*, 14(3):1080–1100, 1986.

[Rusakov and Geiger, 2002] D. Rusakov and D. Geiger. Asymptotic Model Selection for Naive Bayesian Networks. In *Proc. of UAI*, pages 438–445, Alberta, Canada, 2002.

[Sato, 2001] M. Sato. Online Model Selection Based on the Variational Bayes. *Neural Computation*, 13:1649–1681, 2001.

[Schwarz, 1978] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978.

[Stein, 1956] C. Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proc. of the 3rd Berkeley Symp. on Math. Stat. and Prob.*, pages 197–206, 1956.

[Takemura and Kuriki, 1997] A. Takemura and S. Kuriki. Weights of Chi-bar-square Distribution for Smooth or Piecewise Smooth Cone Alternatives. *Annals of Statistics*, 25(6):2368–2387, 1997.

[Wang and Titterington, 2004] B. Wang and D. M. Titterington. Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models with Missing Values. In *Proc. of UAI*, pages 577–584, Banff, Canada, 2004.

[Watanabe and Amari, 2003] S. Watanabe and S. Amari. Learning Coefficients of Layered Models When the True Distribution Mismatches the Singularities. *Neural Computation*, 15:1013–1033, 2003.

[Watanabe and Watanabe, 2006] K. Watanabe and S. Watanabe. Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation. *Journal of Machine Learning Research*, 7:625–644, 2006.

[Watanabe, 1995] S. Watanabe. A Generalized Bayesian Framework for Neural Networks with Singular Fisher Information Matrices. In *Proc. of NOLTA*, volume 2, pages 207–210, 1995.

[Watanabe, 2001a] S. Watanabe. Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, 13(4):899–933, 2001.

[Watanabe, 2001b] S. Watanabe. Algebraic Information Geometry for Learning Machines with Singularities. In *Advances in NIPS*, volume 13, pages 329–336, 2001.

[Watcher, 1978] K. W. Watcher. The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements. *Annals of Probability*, 6:1–18, 1978.

[Yamazaki and Watanabe, 2002] K. Yamazaki and S. Watanabe. Resolution of Singularities in Mixture Models and its Stochastic Complexity. In *Proc. of ICONIP*, pages 1355–1359, Singapore, 2002.

[Yamazaki and Watanabe, 2003a] K. Yamazaki and S. Watanabe. Stochastic Complexities of Hidden Markov Models. In *Proc. of Neural Networks for Signal Processing XIII (NNSP)*, pages 179–188, Toulouse, France, 2003.

[Yamazaki and Watanabe, 2003b] K. Yamazaki and S. Watanabe. Stochastic Complexity of Bayesian Networks. In *Proc. of UAI*, pages 592–599, Acapulco, Mexico, 2003.

[Yamazaki *et al.*, 2005] K. Yamazaki, Kenji Nagata, and S. Watanabe. A New Method of Model Selection Based on Learning Coefficient. In *Proc. of NOLTA*, pages 389–392, Bruge, Belgium, 2005.